Unsupervised Machine Learning Method for Researchers' Profiles Matching

طريقة التعلم الآلي غير الخاضع للإشراف لمطابقة ملفات تعريف الباحثين

Thabit Sulaiman Sabbah
Assistant Professor/ Al-Quds Open University/ Palestine
tazazmeh@qou.edu

ثابت سليمان صبّاح
أستاذ مساعد/ جامعة القدس المفتوحة / فلسطين

Received: 21/09/2021, Accepted: 11/10/2021
DOI: https://doi.org/10.33977/2106-000-005-005
https://journals.qou.edu/index.php/PJTAS

تاريخ الاستلام: 2021/09/21، تاريخ القبول: 2021/10/11
E-ISSN: 2521-411X
P-ISSN: 2520-7431

Abstract

Researcher Profiles Matching is an initial and important step of effective research teams' formation. The researchers' wide, multidisciplinary, and changeable research interests complicate the process of profile matching using traditional methods and affect its performance. This research aims to solve the problem of Profile matching in Scientific Research and Scholarly Work by employing unsupervised machine learning methods. The K-mean clustering method is utilized to categorize researcher profiles based on the statistical analysis of their publication titles, and the correlation-based similarity is employed for profile matching within the categories. The proposed method is implemented, tested, and evaluated using an extracted dataset from Google Scholar. The profile matching results and the clustering quality test result show that the designed task was achieved, in addition to high similarity values of publications within the categories and low correlation values among the clusters. Moreover, the clustering results' analysis can reveal interesting and enlightening information about the scholarly work, which may help the researchers, research management departments, as well as policies and decision-makers in their scholarly work associated tasks.

***Keywords**: Researcher Profiles Matching, Unsupervised Machine Learning, Correlation-based Similarity, K-mean algorithm, Google Scholar.*

الملخص

مطابقة ملفات تعريف الباحثين هي خطوة أولية ومهمة لتشكيل الفرق البحثية الفعالة. إن الاهتمامات البحثية الواسعة ومتعددة التخصصات والمتغيرة للباحثين تُعَقِّد عملية مطابقة الملفات التعريفية باستخدام الأساليب التقليدية، وتؤثر على أدائها. يهدف هذا البحث إلى حل مشكلة مطابقة الملفات الشخصية في مجال البحث العملي، والعمل البحثي من خلال توظيف طرق تعلم الآلة غير الخاضعة للإشراف. واستخدمت طريقة التصنيف (ك-متوسطات) لتصنيف ملفات تعريف الباحثين اعتمادا على التحليل الإحصائي لعناوين أبحاثهم، ووظِف التشابه المبني على الارتباط لمطابقة ملفات التعريف ضمن الفئات. وتم بناء الطريقة المقترحة، وفحصها، ثم قُيِّمت باستخدام مجموعة بيانات مستخلصة من موقع الباحث العلمي ل(جوجل). وأظهرت نتائج مطابقة الملفات الشخصية، وفحص جودة التصنيف أن المهمة المصممة قد تمّ إنجازها، يضاف إلى ذلك ظهور قيم تشابه عالية للأبحاث داخل الفئة وقيم ارتباط متدنية بين الفئات. ويمكن لتحليل نتائج التصنيف أن تكشف معلومات مضيئةً ومهمةً حول العمل البحثي، والتي من شأنها أن تساعد الباحثين، ودوائر إدارة البحث، وصُناع السياسات والقرارات في مهامهم المرتبطة بالعمل البحثي.

الكلمات المفتاحية: مطابقة ملفات تعريف الباحثين، تعلم الآلة غير الخاضع للإشراف، التشابه المعتمد على الارتباط، خوارزمية ك-متوسطات، الباحث العلمي.

INTRODUCTION

Researcher profiles matching is a special case of the general known problem of User Profile matching, which has been tackled in several works over the years. It is a part of the team formation process encouraged by mast organizations to carry out complex tasks (Sun et al., 2009). Many other profits and benefits can be brought to the organization because of effective teams. However, the environment in which the team will be formulated, the task to be accomplished, and many other factors affect the formation process and criticality. Some of these factors are related to the team size, distribution (Milojević, 2014), available data about users (Nurgaliev et al., 2020), and such as the case of team formulation in complex networks and large communities (Sun et al., 2009). On the other hand, from individuals' (researchers) perspectives, researcher profile matching helps in finding potential research collaborators, expertized researchers in a certain domain, expanding network opportunities (Tran et al., 2020), and improving profile building skills (Li et al., 2019).

A research team is defined as a "group of researchers collaborating to produce scientific results, which are primarily communicated in the form of research articles" (Milojević, 2014). A research team may consist of some core researchers and many other researchers who may change over time. Hence there are many works focused on studying the statistical measures of a team such as size, median, and mean, assuming that teams are unchangeable, while fewer studies consider the changeability of teams (Milojević, 2014).

Therefore, several models are proposed in these studies for different cases, aims, bases, and domains such as the Agent-based model (Sun et al., 2009), supervised ML model (Nurgaliev et al., 2020), and others. This work presents the unsupervised machine learning clustering method for researcher profile matching based on researchers' publications metadata available on Google Scholar, such as researcher interests, and publication titles. The rest of this article contains sections about related works, proposed method, methodology, results discussions, and conclusion.

LITERATURE REVIEW

As mentioned earlier, Profiles Matching was a well-known problem that was studied from different aspects over the years in many domains. However, fewer studies were found in the domain of Scientific Research (i.e., matching researchers' profiles to find potential research collaborators and expertized researchers in a joint domain). Therefore, this section summarized the existing works on profile matching and the unsupervised machine learning clustering method utilized by this study.

Profile Matching Works

Profile Matching Algorithm (PMA) was employed in many fields such as business, social networks, and others, following a brief summarization of some studies from different domains.

In the business domain, Sugiarto et al. (2021) described the use of PMA in the context of a decision support system that could help shorten the required time for choosing business partners or potential colleagues in companies. However, the study focused on analyzing input factors of the PMA and the GAP calculations and weightings. The study concluded that the application of PMA based on predetermined conditions could accelerate model calculation and select prospective partners' processes.

Nurgaliev et al. (2020) proposed PMA that dealt with a set of linked nodes from various social networks based on inadequate user profile data such as username and relationship. The proposed framework included two individual algorithms and a combination of them. The proposed User identity linkage (UIL) algorithm aimed to determine mathematically whether any two users on different social networks are the same person in reality. The proposed algorithms were tested on datasets from VK social network and Instagram; the experiments showed relatively high recall and accuracy results.

Eze et al. (2020) presented a configurable PMA in the domain of health community care management. The work aimed to associate common data from various stakeholders to support the process in the domain. Eze et al. (2020) focused on the performance of PMA utilization in the cloud-hosted case study. They tested the proposed model within a pilot project for supporting interoperability between Community Support Service (CSS) provider agencies and the Regional Health Authority (RHA) in Canada. The Proposed PMA consisted of many modules such as feature identification, standardization, match weight summarization, decision, and global identifier generation. The first run of the system was conducted based on about 145,000 user-profiles and took about 35 minutes; however, the sequent daily runs performed the task incrementally and required less than 5 minutes per day.

Similarly, Li et al. (2019) applied the PMA to find the match users' profiles under the condition of restricted data access of users' profile data such as profiles with privacy policies. The proposed method in Li et al. (2019) utilized the public data such as username and display name and accomplished the matching task through a three-step approach, including feature extraction, a two-stage classification framework, and a relationship elimination algorithm. Experimental results on real social networks datasets showed excellent performance and concluded the possibility of applying PMA based on small and public online user profile data.

Paembonan et al. (2018) employed the PMA for drug substitution to facilitate the process of drug substitution in cases of drug lack or exhaustion. The K-means method was utilized to categorize the medicines' profiles to accomplish the task of new medicine recommendations, where the Selection Matching method was employed to control the substitute. The proposed method was tested and evaluated. The authors reported the accuracy of the proposed method was 93.5%.

Earlier, many works have been presented in the field of User Profile Matching, such as (Garcia, 2016; Pizzi and Ukkonen, 2008; Sun et al., 2009; Wassermann and Zimmermann, 2011).

Nevertheless, none of these works was in the field of Scientific Research or Researchers Profile matching and applying any unsupervised machine learning clustering techniques. Although the work of (Paembonan et al., 2018) utilized the k-means algorithm, the work does not explain much about utilizing K-means with PMA. Therefore, this work tried to accomplish the process of PM in the field of Scientific Research by applying some unsupervised machine learning clustering techniques. The following subsection illustrated the principles of unsupervised clustering methods and described the k-means clustering method.

Unsupervised Clustering Methods

Clustering was defined as "the unsupervised classification of data objects into groups or clusters" (Santos et al., 2013). The term "unsupervised" indicated that the process was done under the condition of missing ground-truth labels of classified objects. Therefore, unsupervised clustering methods must first notice any patterns in the data objects being clustered and then group similar objects in a category such that the objects in a group were the most similar to each other. This process of clustering was unlike supervised learning (known as supervised classification), where human experts usually provided the ground-truth labels of the training data. These unsupervised clustering advantages were included but not limited to a slight workload to audit and formulate training data, and superior independence in identifying and utilizing hidden patterns that "experts" had not observed. However, the cost of such benefits included the need for more amount of data for training to achieve acceptable performance which indicated extra storage and computational necessities, as well as the possibility of such method to consider some anomalies or artifacts found in training data as bases of clustering (Delua, 2021). Many methods and techniques were used for clustering such as hierarchical clustering (Franklin, 2005), and k-means which was one of the popular and simplest unsupervised machine learning algorithms (Garbade, 2018).

K-Means Clustering Algorithm

Andrews and Fox (2007) considered this algorithm as the most regular and simple algorithm used for clustering. The algorithm aimed to group the nearest data objects to each other onto smaller sets. A key point for the algorithm was the determination of the number of clusters. After this determination, the algorithm spread the data objects into the determined number of clusters based on objects' features, reflecting the likeness of the data objects (Jain et al., 1999). As mentioned earlier, this clustering method was employed in many fields such as "Topic Detection." For example, Li et al. (2010) performed a study in which the k-means algorithm was employed on top of the Vector Space Model (VSM) representation to detect topics among a corpus. Similarly, Zhang and Li (2011) proposed the k-means clustering method for topic detection in a large-scale dataset. The K-means algorithm was performed by applying the following steps:

1. Determine the number of clusters (the value of k).
2. Randomly select k data objects as preliminary cluster centers (in some implementations, the first K data objects were selected for this step).
3. Calculate the *distance* between the defined cluster centers and the remaining data objects, and assign each data object to a cluster center based on the nearness of the cluster center.
4. For each defined cluster, calculate the mean and update the cluster center to become the calculated mean.
5. If no change occurred to any cluster center values, then STOP, otherwise repeat steps 3-5.

Nevertheless, the k-means clustering method had some downsides, such as its sensitivity to the initial selection of cluster centers, as well as its sensitivity to outliers and noise, and the non-predefined number of clusters. These drawbacks might constitute inaccuracy (Sharma and Gupta, 2012) or unwanted solutions (Jain et al., 1999). However, several techniques were proposed in the literature to overcome these problems. For example, Ray and Turi (1999) recommended the validity measure to determine the k number. Some other works were planned to solve the problem of finding the preliminary cluster centers using different principles, such as Erisoglu et al. (2011), Deelers and Auwatanamongkol (2007), and Redmond and Heneghan (2007).

The distance calculation mentioned in step 3 of the k-mean algorithm differed according to the domain of application. For example, in case that

the data points to be clustered were point 2D or 3D Cartesian coordinate system, the regular distance law between points in such coordinate system and be applied. However, when applying the k-means algorithm in other domains, such as text clustering where the data points represent the documents, the Euclidian distance or the Cosine similarity could be applied. In this research, the algorithm was applied to multi-dimensional feature space. Therefore, the Euclidian Distance Law was applied. The Euclidian distance between two documents represented in a high dimensional feature space was defined as follows:

Let the two data points (i.e., documents) to be A and B, where A and B were vectors of n features such that: A= $\{a_0, a_1, a_2, \ldots, a_n\}$ and B = $\{b_0, b_1, b_2, \ldots, b_n\}$, then the Euclidian distance D between these two data points was calculated according to equation (1).

$$D(A,B) = \sqrt{\sum_{i=0}^{n}(a_i - b_i)^2} \quad (1)$$

The next section explained the proposed method for Researcher Profile Matching.

MATERIALS AND METHODS

This work proposed an Unsupervised Machine Learning Clustering Method for Researcher Profile Matching. The proposed method was based on the analysis of user-profiles data from Google Scholar (GS) Search Engine. A researcher profile on GS contained many informative data portions such as interests, count and distribution of publications over the years, h-index, i10th index, count of citations, and publications list. Nevertheless, some of these elements might be missing or incomplete or not updated in some user profiles. Therefore, some of these elements were utilized in this work for profile matching, especially the publication list, which reflected researcher interests. The next subsections showed the details of the proposed matching method and the dataset used in this work.

Proposed Matching Method

Figure 1 demonstrates the proposed method steps and processes followed by a brief description of the shown steps, where each numbered bounded area was considered as one step, and the method consists of five steps.

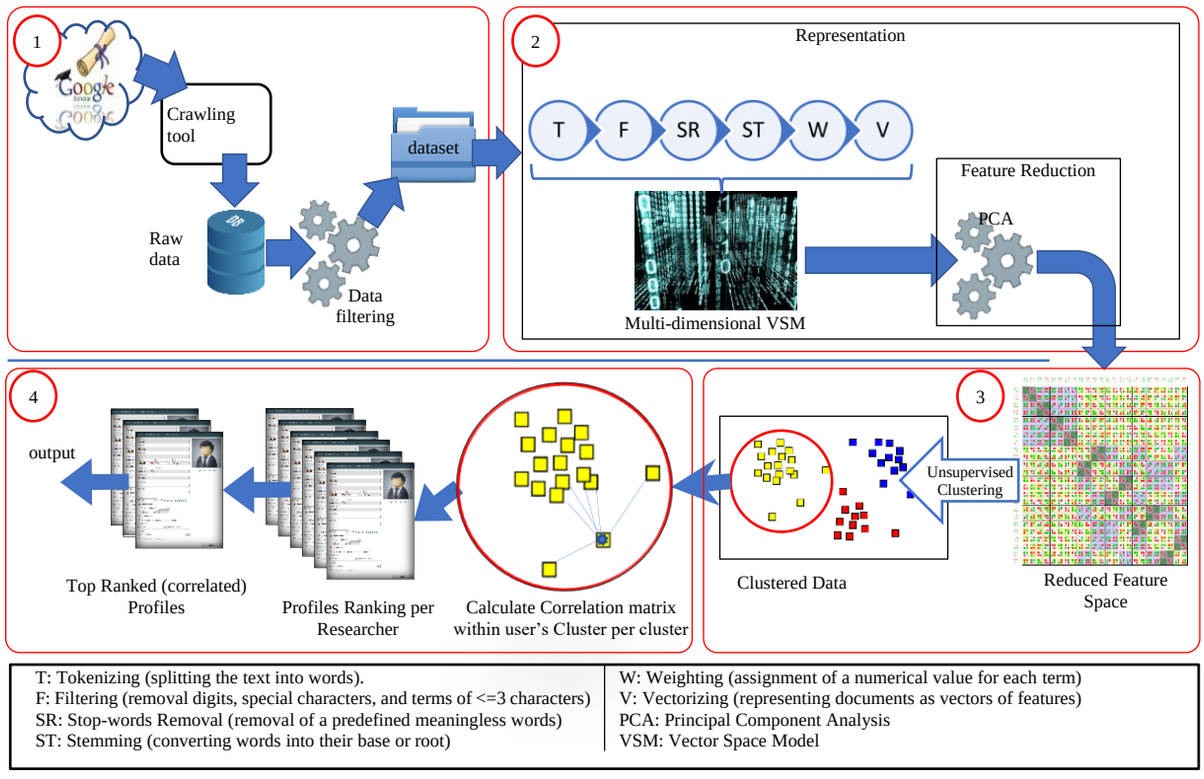


Figure 1 Proposed Researcher Profile Matching Method Steps and Processes

Step 1: This step was devoted to Dataset generation. Dataset description was shown later in the section. In this step, a crawling tool was developed to download hundreds of researcher profiles from GS. These profiles were stored on a local database in HTML format, and then it was processed, filtered, and prepared as the final dataset. The researcher profile on GS contained many portions of data; the distribution of these data chunks on the researcher's profile page was as shown in Figure 2.

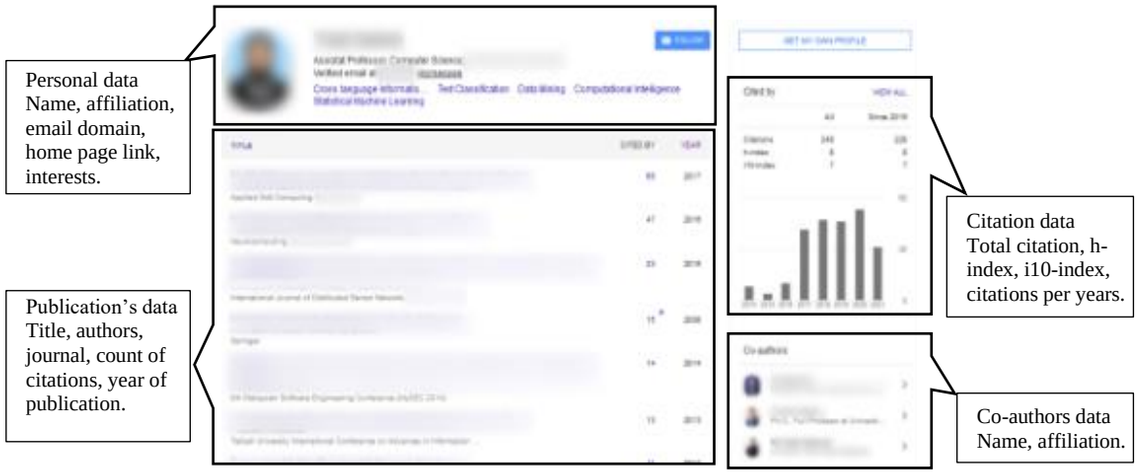


Figure 2 Distribution and Data Chunks on Researcher's Profile Page on GS

However, as mentioned earlier, some of these data chunks might be missing, incomplete, or not updated in some user profiles. Moreover, some of the researcher's publication lists might contain multi-lingual titles. The list consisted of hundreds of publications. Therefore, this step included a filtering process for the publications within the last five years, in which the titles in the English language were considered in the dataset.

Step 2: The dataset was presented numerically to be suitable for the Machine learning methods. The Vector Space Model (VSM) representation was considered in this work. A series of preprocesses tasks were performed for each textual data for each instance in the dataset to achieve this representation. These tasks are: ***Tokenizing, Filtering, Stop-words Removal, Stemming, Weighting, and Vectorizing***. A brief description of these tasks is presented at the bottom of Figure 1. However, regarding **Weighting**, which was the process of assigning a numerical value for each word (term or feature) per dataset instance. This numerical value of a term (known as term weight) represented the importance of that term in that instance. In literature, there were many weighting techniques such as the binary, the Term Frequency (TF), the Term Frequency-Inverse Document Frequency (TF-IDF), and many more (Sabbah et al., 2017). However, this work utilized the Term Occurrence (TO) method that considered the count of term appearance as the term weight. This technique of term weighting, i.e., TO did not consider the normalization of weighing such as the TF and TF-IDF techniques; moreover, it did not show any kind of semantic proximity such as the Term Co-occurrence weighting method. The choice of TO weighting technique in this research was based on the nature of the processed text (i.e., Publication Titles), which was assumed to be clear, specific, and direct to the point.

Vectorizing: In this process, each data sample was represented as a vector of features, where the features of the vector included all the features (terms) contained by the dataset. The vectors were finally collected in one matrix. The rows represented the data samples, the columns represented the features, and the matrix's cells' values represented the weights.

Feature Reduction

The generated VSM based on text vectorization was known as multi-dimensional, in which the count of features was large. For example, during our experiments, the count of features based on the unigram vectorization of publication titles and publication summaries was more than 450,000 features, i.e., unique single word, which was out of our capability to

manipulate due to lack of computational capacity). Therefore, we restricted the textual analysis in this work to publication titles where the count of features in the generated feature space was about 25000 features, which was huge. Therefore, the Principal Component Analyses (PCA) dimensionality reduction method was employed to reduce the dimensionality, reducing computational cost and time.

Step 3: K-means clustering - which was an unsupervised machine method- was a learning method applied to categorize the data samples into clusters or categories where the categories represented the research fields or research topics reflected from publication titles. However, there was a wide range of research fields or topics that could be identified. Thus, the determination of clusters count- that represented the K value in the K-means algorithm- was not an easy task. To do so, the lists of research fields were studied from different sources, as follows:

Table 1 Count of Research Fields from Different Online Sources

List Source	Count of Research fields
Wikipedia: (https://en.wikipedia.org/wiki/Outline_of_academic_disciplines)	1000
Digital Commons Network™: (*https://network.bepress.com*)	1280
Web of Science (WoS): (https://images.webofknowledge.com/images/help/WOS/contents.html)	258
Higher Education Statistics Agency (HESA), UK: (*https://www.hesa.ac.uk/support/documentation/jacs*)	165
Japan Society for the Promotion of Science (JSPS), Japan: (*https://www.jsps.go.jp/english/index.html*)	323

Table 1 showed that the count of research fields was not standard and differed from one source to another, and the count was not enclosed in a small range. Therefore, it was a challenge to determine the count of research fields (i.e., clusters). Nevertheless, there were several computational based techniques to automatically determine the best value of (K), such as the Distortion Analysis (known as Elbow Curve Method) (Yuan and Yang, 2019), Davies-Bouldin Index (Petrovic, 2006), and Calinski-Harabasz Index (Wang and Xu, 2019) and more. Hence, in this study, the results of these techniques were analyzed to determine the best value of K (i.e., clusters count). However, the application of these methods was time-consuming, as the algorithm was required to run numerous times based on various values of K for each technique, which was applicable only for small datasets and K values. However, in our case, the potential value of K was as high as expected by the common sense shown in Table 1, and the dataset size was as big as shown in the dataset subsection. Hence, a sample dataset selected from the study dataset was employed for exploratory study and determination of the count of clusters (i.e., K value for K-means algorithm). The details of the exploratory dataset and the K value determination analysis was shown in the next subsections.

Moreover, Step 3 produced the cluster label for each sample in the dataset. Consequently, these labels were utilized in Step 4 for profile matching.

Step 4: In this step, for each cluster of the identified clusters, the samples that belonged to that cluster were identified and isolated, and then the correlation-based similarity was calculated among all samples within the cluster, the samples such as profiles were ranked, and the top similar correlated profiles were recommended as the best matching profiles for any selected user.

Dataset

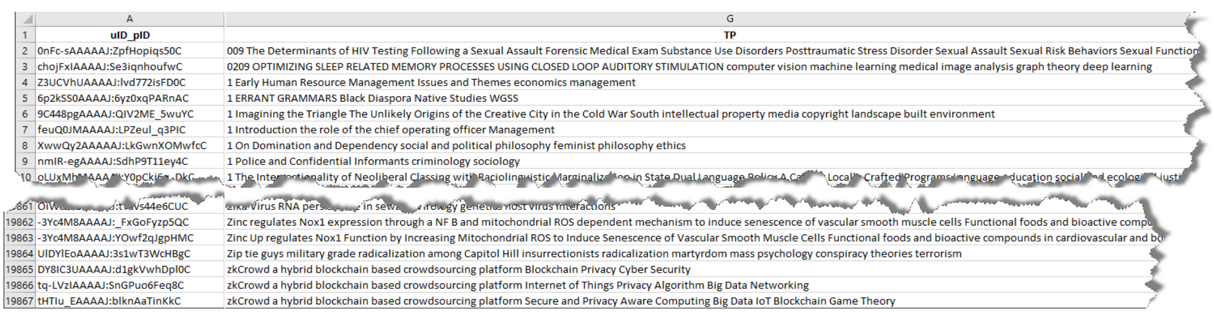
As mentioned in Step 1 description, hundreds of researcher profiles were crawled from GS as Html web pages. The data chunks were extracted from the web pages and filtered. The data chunks that could be utilized are many, such as Researcher's Years of Experience (RYE), h_Index (hI), i10_Index (iI), Publication Age (PA), Publication Citations Count (PCC), Publication Title (PT), and Researcher List of Interests. In addition to the user ID and publication ID (uID:pID) for indexing and matching purposes. However, some of these chunks of data were user-related, such as RYE, hI, iI, and research interests, while others were publication-based, such as PA, PCC, and PT. Therefore, as this study focuses on textual-based categorization and profile matching, the publication-based chunks of data were considered in the dataset, especially the Publication Title (PT). Nevertheless, Regarding the Researcher List of Interests, it was noticed during preprocessing that the keywords included in the List of Interests of researchers contained noise data such as spelling mistakes and

sometimes not. Therefore, the interests' keywords were treated in this work as "text" and added to each publication's title found in the researcher profile. Initially, the dataset represented the data of 1351 researchers from Georgia State University (GSU) and included the data of 22540 publications. However, as a part of data filtering mentioned in Step 1, the Researcher Profiles, which included a very large or very low count of publications (outliers), were eliminated; for experimental purposes, the elimination was performed using a simple query method. The remaining profiles contained a count of publications ranging from 6 up to 2239 publications. Table 2 shows the statistical information of the final dataset. Figure 3 shows a portion of records in CSV format, while Figure 4 and Figure 5 show the most frequent words and Bigrams used in the publication Titles included in the final dataset.

Table 2 Dataset Statistics

Count of Users (Researches)	882	
Count of Publications	19866	
Publication Titles		
Unique Vocabulary in Publication Titles	18350	
Total count of words in publication Titles	269854	
Average words count per Title	14.94	
Publication Title Statistics	Publications Count	Title Length (words)
Average	44.99	143.01
Min.	6	10
Max.	239	312

As seen in Table 2, the final dataset contained the data of 882 Researchers and included the titles of 19866 publications. The eliminated profiles, i.e., the profiles which included a very large or very low count of publications, perform about 34% of the total profiles count. However, the effect of this elimination in terms of computational cost, performance, and time was not studied in this research as this research aimed to prove the concept of the proposed method.



	A	G
1	uID_pID	TP
2	0nFc-sAAAAAJ:ZpfHopiqs50C	009 The Determinants of HIV Testing Following a Sexual Assault Forensic Medical Exam Substance Use Disorders Posttraumatic Stress Disorder Sexual Assault Sexual Risk Behaviors Sexual Function
3	chojFxIAAAAJ:Se3iqnhoufwC	0209 OPTIMIZING SLEEP RELATED MEMORY PROCESSES USING CLOSED LOOP AUDITORY STIMULATION computer vision machine learning medical image analysis graph theory deep learning
4	Z3UCVhUAAAAJ:lvd772isFD0C	1 Early Human Resource Management Issues and Themes economics management
5	6p2kSS0AAAAJ:6yz0xqPARnAC	1 ERRANT GRAMMARS Black Diaspora Native Studies WGSS
6	9C448pgAAAAJ:QIV2ME_5wuYC	1 Imagining the Triangle The Unlikely Origins of the Creative City in the Cold War South intellectual property media copyright landscape built environment
7	feuQ0JMAAAAJ:LPZeul_q3PIC	1 Introduction the role of the chief operating officer Management
8	XwwQy2AAAAAJ:LkGwnXOMwfcC	1 On Domination and Dependency social and political philosophy feminist philosophy ethics
9	nmIR-egAAAAJ:SdhP9T11ey4C	1 Police and Confidential Informants criminology sociology
10	[illegible]	1 The [illegible] of Neoliberal Classing with Raciolinguistic [illegible] in State Dual Language Policy [illegible]
[illegible]	[illegible]	[illegible] Virus RNA [illegible] host virus interactions
19862	-3Yc4M8AAAAJ:_FxGoFyzp5QC	Zinc regulates Nox1 expression through a NF B and mitochondrial ROS dependent mechanism to induce senescence of vascular smooth muscle cells Functional foods and bioactive compo
19863	-3Yc4M8AAAAJ:YOwf2qJgpHMC	Zinc Up regulates Nox1 Function by Increasing Mitochondrial ROS to Induce Senescence of Vascular Smooth Muscle Cells Functional foods and bioactive compounds in cardiovascular and bo
19864	UlDYlEoAAAAJ:3s1wT3WcHBgC	Zip tie guys military grade radicalization among Capitol Hill insurrectionists radicalization martyrdom mass psychology conspiracy theories terrorism
19865	DY8IC3UAAAAJ:d1gkVwhDpl0C	zkCrowd a hybrid blockchain based crowdsourcing platform Blockchain Privacy Cyber Security
19866	tq-LVzIAAAAJ:SnGPuo6Feq8C	zkCrowd a hybrid blockchain based crowdsourcing platform Internet of Things Privacy Algorithm Big Data Networking
19867	tHTIu_EAAAAJ:blknAaTinKkC	zkCrowd a hybrid blockchain based crowdsourcing platform Secure and Privacy Aware Computing Big Data IoT Blockchain Game Theory

Figure 3 A Snap of Dataset in CSV Format

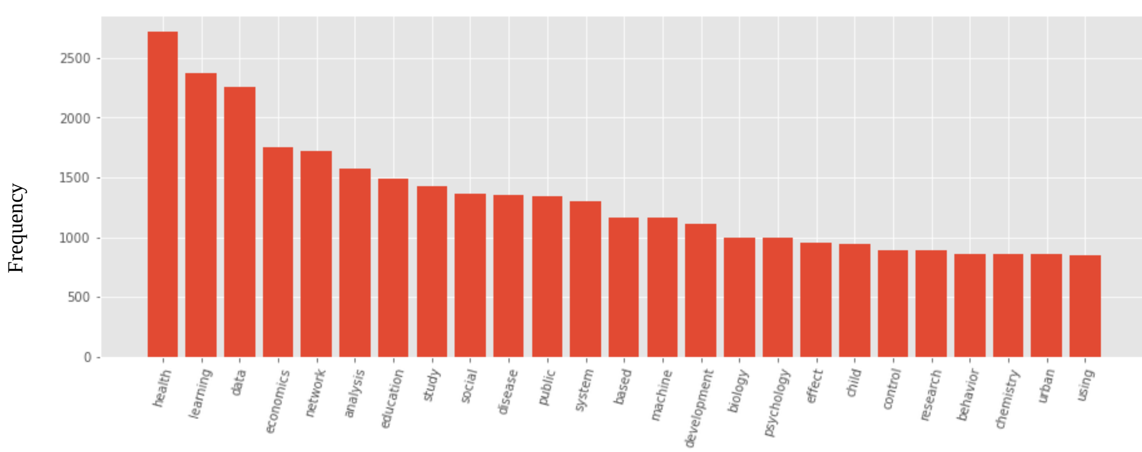


Figure 4 Most Frequent Words used in Publication Titles

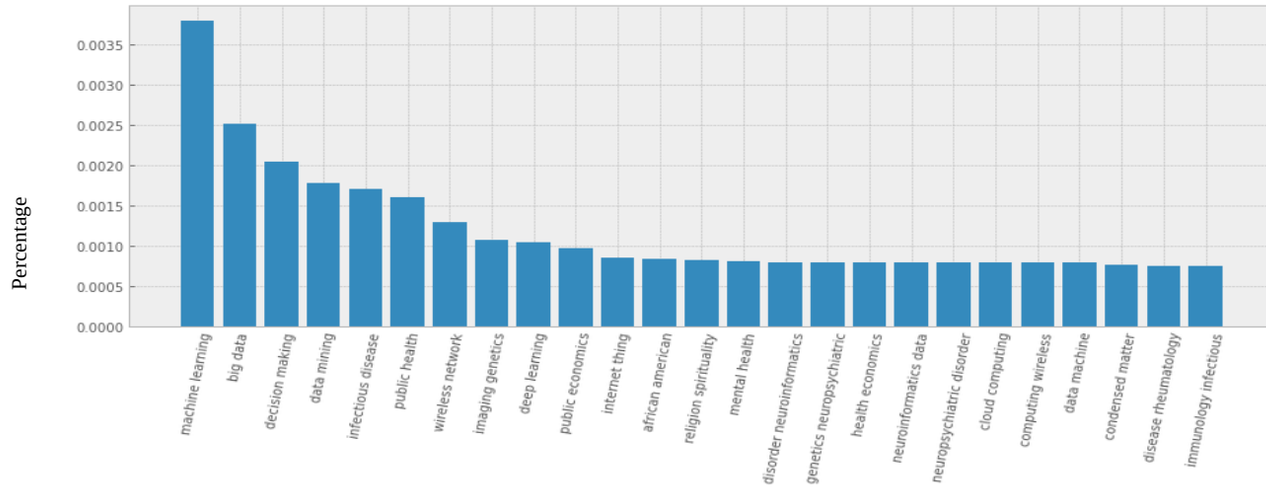


Figure 5 Most Frequent Bigrams used in Publication Titles

As mentioned in Step 2, the PCA method was utilized to reduce the size of feature space generated by the VSM. The application of the PCA algorithm in this step reduced the size of the feature space by about 78%, such that the number of features was reduced from 18350 features in the VSM feature space to 4217 features in the reduced feature space. However, the size of the reduced feature space was selected to represent about 95% of the original feature space. Figure 6 shows the size of reduced feature space after PCA application.

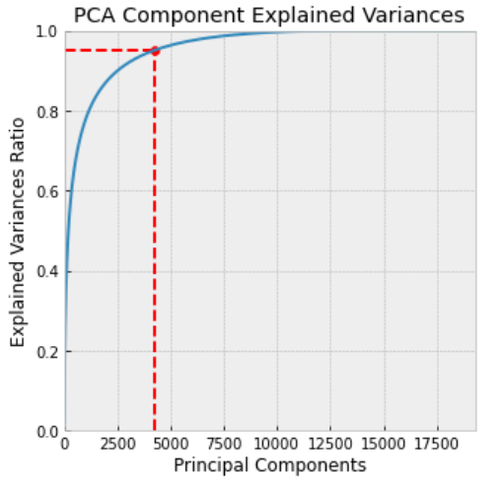


Figure 6 Size of The Reduced Feature Space after PCA Application

As seen from Figure 6, initially, the count of principal components was equal to the count of features in the VSM feature space. However, the curve showed the cumulative variance explained by these components. The cumulative variance of the selected count of components, i.e., 4217, explained 95% of the feature space.

Exploratory Dataset

As the Step 3 discussion mentioned, the researcher employed an exploratory dataset to determine the K-value required for the K-means algorithm. Then selected the exploratory dataset to be representative and informative. Therefore, for each researcher among the 882 researchers included in the study dataset, three publications were selected so that the top three cited publications were included in the exploratory dataset. The selected publications per researcher (i.e., top-cited publications) were expected to be the nearest (or representing the field of study of the researcher). Table 3 showed the statistics of the exploratory dataset. Figure 7 and Figure 8 showed the most frequent words and Bigrams used in the publication Titles included in the exploratory dataset.

Table 3 Exploratory Dataset Statistics

Count of Users (Researches)		882
Count of Publications		2646
Publication Titles		
Unique Vocabulary in Publication Titles		7153
Total count of words in publication Titles		38723
Average words count per Title		14.63
Publication Title Statistics	**Publications Count**	**Title Length (words)**
Average	3	14.63
Min.	3	1
Max.	3	36

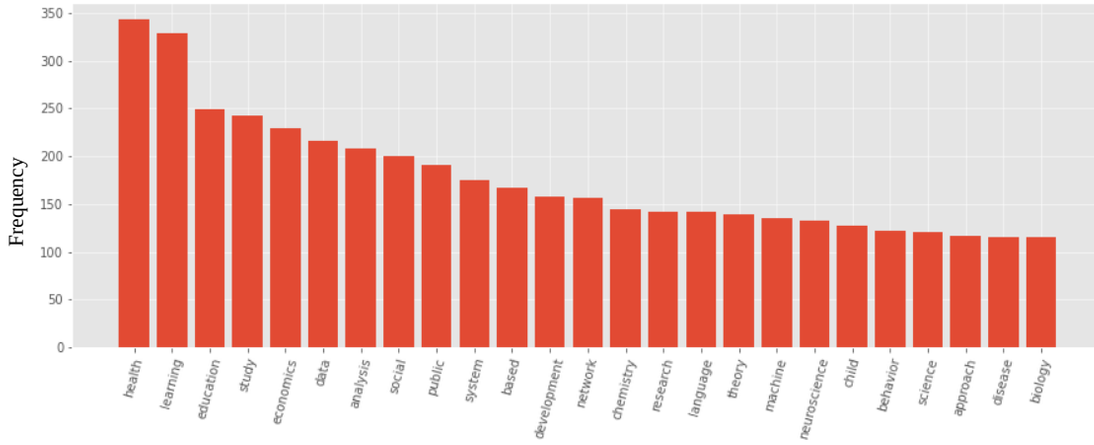


Figure 7 Most Frequent Words used in Publication Titles in The Exploratory Dataset

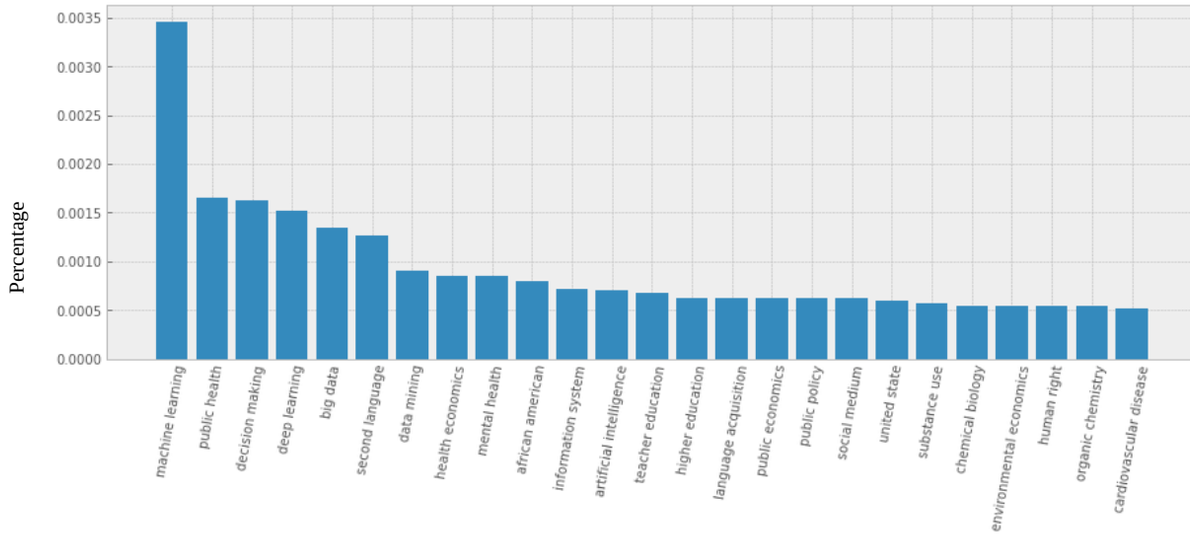


Figure 8 Most Frequent Bigrams used in Publication Titles in The Exploratory Dataset

RESULTS AND DISCUSSIONS

The proposed method was implemented in Python 3.8, while the experiments were conducted under Windows 10 environment, and the results were analyzed using a collection of tools including Orange and MS-Excel. Tasks of clustering and matching were accomplished.

However, as it was known about clustering methods, the evaluation of the clustering was as difficult as the clustering itself (Pfitzner et al., 2008). The proposed method in this work tried to solve the problem of Profile Matching through the application of clustering, an unsupervised machine learning method. Nevertheless, none of the problems - i.e., the profile matching and the clustering- in the domain of consideration had a gold standard dataset to evaluate the results of the proposed method. Therefore, the internal method of evaluation (Feldman and Sanger, 2006) was applied in which the internal clustering quality measures were analyzed, then the corresponding results of the proposed method were benchmarked. Recall that the scope of this work did not include Topic detection, i.e., the method was not responsible for knowing the Research Field of a researcher or publication. However, the clusters or categories in this research represented the Research Fields. Therefore, in this section the clusters were presented by their given numbers: 0, 1 … and so on. Following are the major finding based on the analysis of the results.

As mentioned earlier, the count of clusters considered in this work was determined based on the analysis of the three different clustering quality techniques results on an exploratory dataset; the next subsection shows this analysis' results.

K-value Determination

To determine the optimal value of K, the algorithm was run with K value ranges from 100

to 310. The clustering quality measures; Distortion, the base of Elbow Curve analysis, Davies-Bouldin Index, and Calinski-Harabasz Index, were recorded, scaled, and plotted as shown in Figure 9.

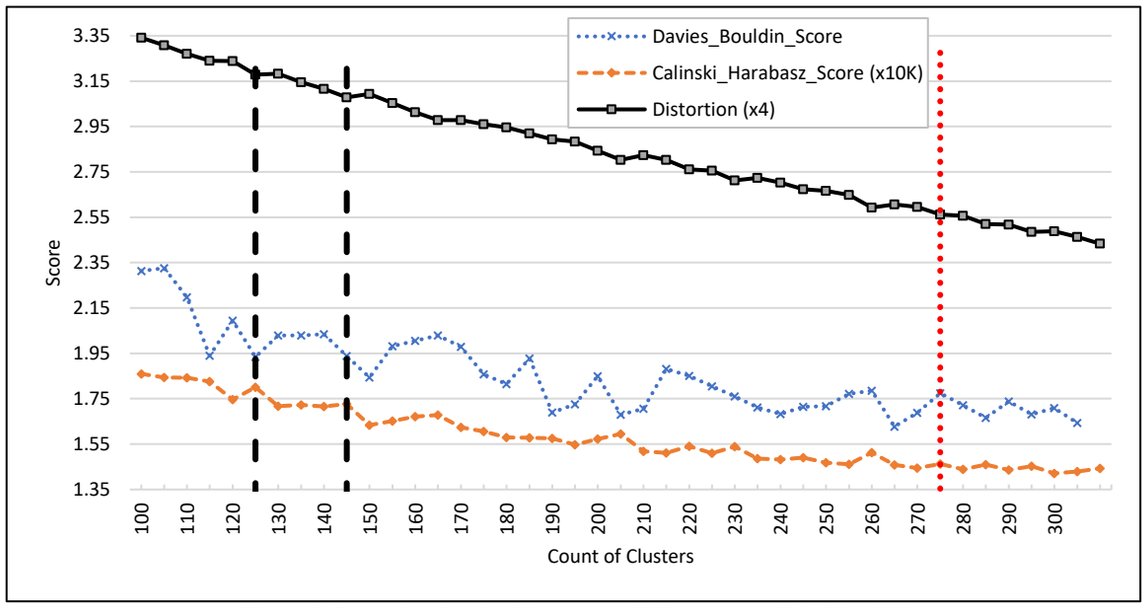


Figure 9 Clustering Quality Measures of K Values from K=100 to 310

Figure 9 showed that several potential K values produce satisfactory quality measures and can be considered as the count of clusters. Based on the assumptions behind these three measures, the value K = 275 was selected as the cluster count in this research. Moreover, a further Kolmogorov Smirnov test (Wilcox, 2017) at that value of K, i.e., 275 showed that the distribution of samples among the clusters fit the normal distribution with a value of $p < 0.05$. The next subsection presented the analysis of results from two-point views: Researchers' Distributions and Publications Distribution against Research Fields, i.e., Categories or Clusters.

Researchers Distributions Analysis

Figure 10 shows the distribution of the count of Researchers among these clusters.

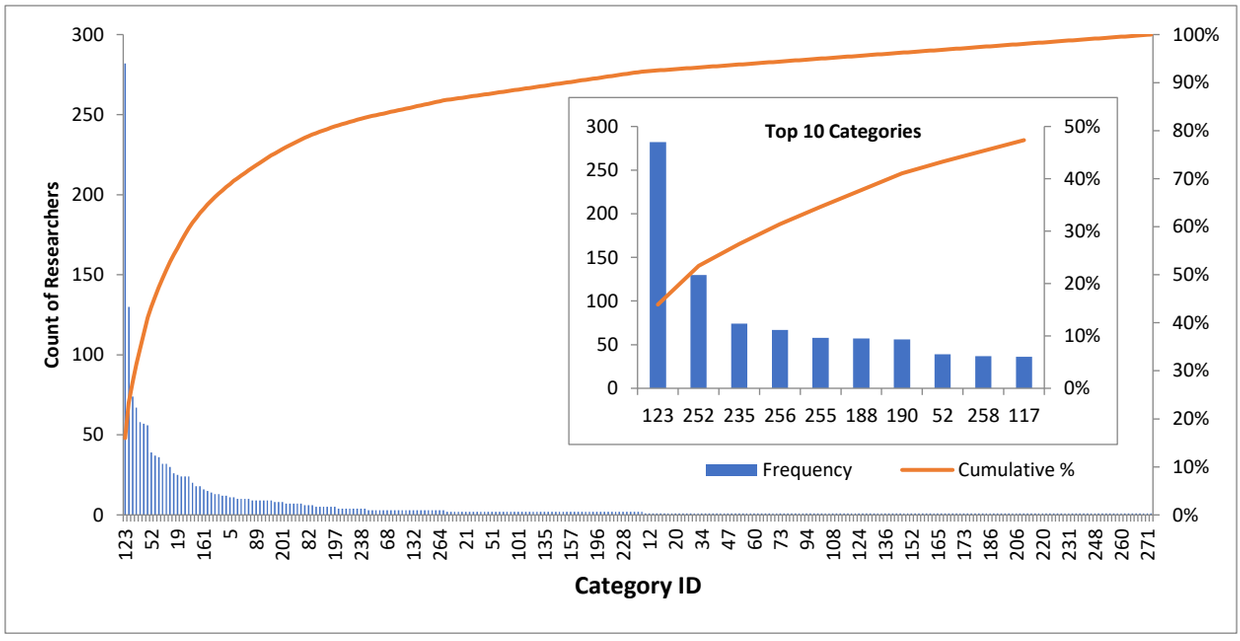


Figure 10 Researchers Distribution among Clusters (Research Fields)

Figure 10 shows the distribution of researchers among the 275 considered clusters. The Frequency columns represented the count of researchers belonging to the corresponding cluster from the horizontal axis. The distribution of the researchers among the "Top 10 Categories" was shown in the internal subplot ("Top 10 Categories"). For example, there were about 282 researchers grouped in "Cluster 123", while less than half of this count of researchers 130 in "Cluster 252", for the remaining clusters, the count of researchers ranges between 1 and 75 researchers. Moreover, the "Top 10 Clusters" included about 50% of the researchers' distribution, whereas the 9th and 10th clusters contained less than 50 researchers each. It is worth mentioning that researchers' categorization (clustering) was based on their scholarly production within the last five years. Therefore, some (if not many) researchers were identified to be included in multiple clusters, which reflected the multidisciplinary nature of many researchers. Figure 11 shows the multidisciplinary distribution identified in the considered dataset.

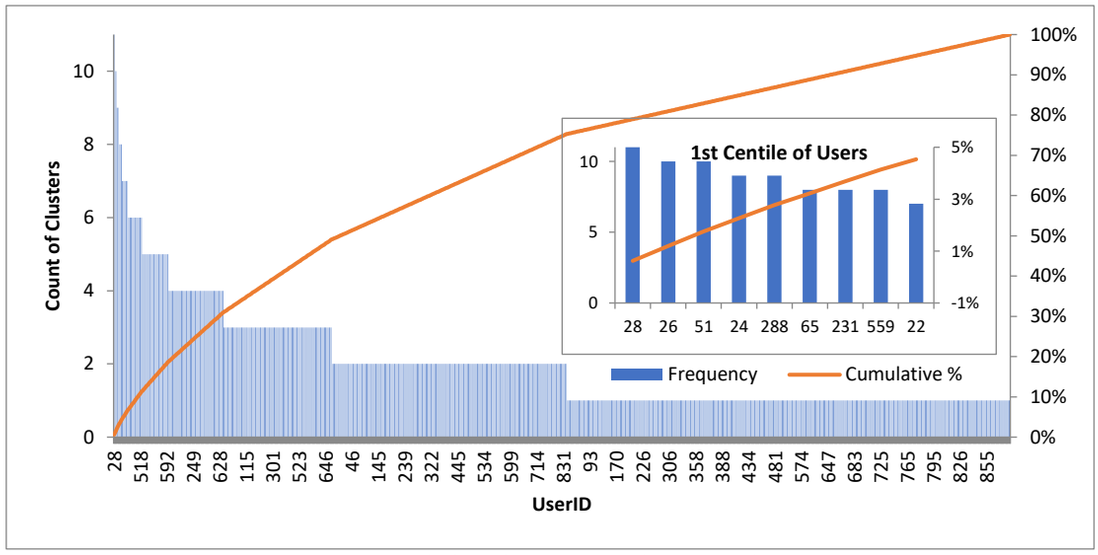


Figure 11 Multidisciplinary Researchers' Distribution

From the multidisciplinary distribution of researchers shown in Figure 11, very few researchers were categorized as involved in abundant research fields; more than 7 fields as revealed in the inner subplot, i.e., the 1st centile of users' multidisciplinary distribution. This portion could be caused by outlier profiles in which a huge number of publications were added automatically to a researcher profile because of the known problem of initials ambiguity of researcher names (Milojević, 2013). Figure 12 shows the clusters per user distribution.

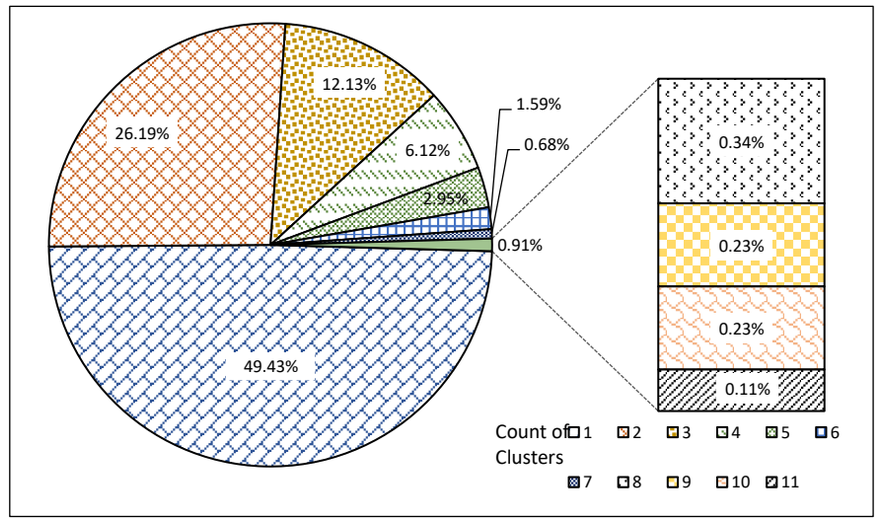


Figure 12 Researchers Distribution among Research Fields

Figure 12 showed that the majority, about 87.5% of researchers, were identified to be working within limited research fields at most 3 disciplines. 49.43% were mono disciplinary researchers, 26.19% were involved in two disciplines, and 12.13% were contributary to three research disciplines. About 11% of researchers were identified to be involved in -4 to 7- research fields and the remaining less than 1% of researchers were involved in abundant research fields as described earlier.

Publications Distribution Analysis

Figure 13 showed the publications distribution among the research fields.

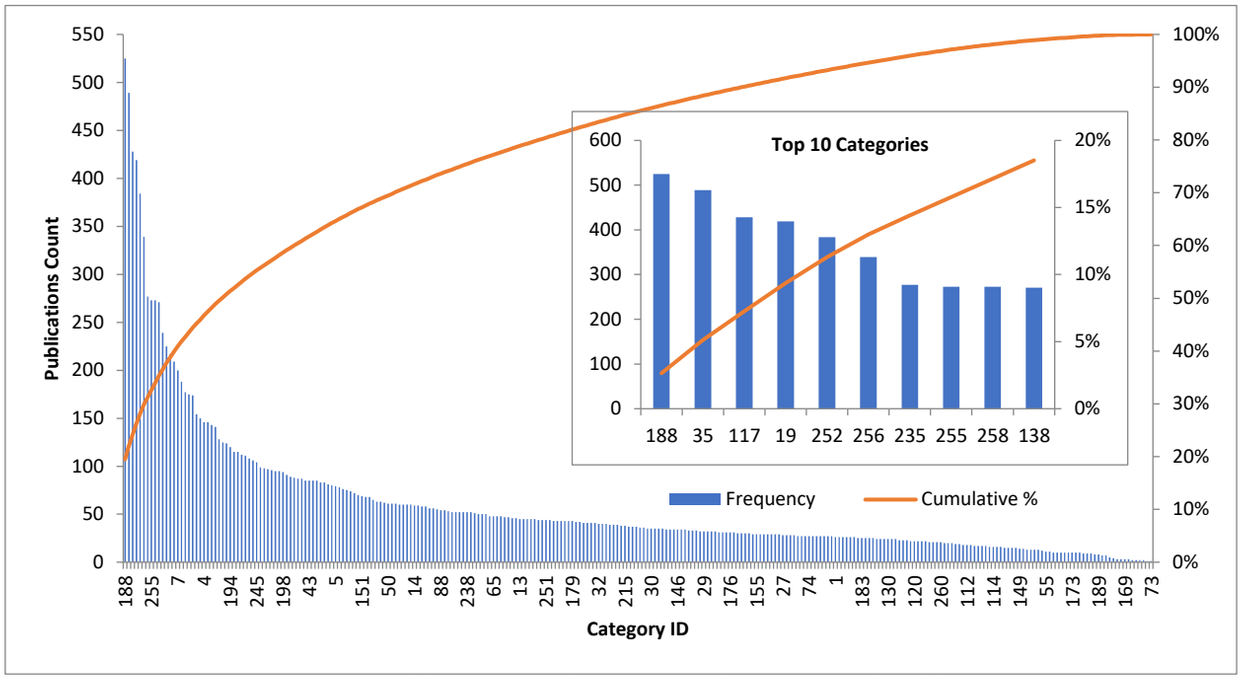


Figure 13 Publications Distribution among Clusters (Research Fields)

Figure 13 shows the distribution of Publications among the 275 Research fields considered in this study. The Frequency columns represented the count of publications that belonged to the corresponding cluster from the horizontal axis. The distribution of publications among the top 10 Categories was shown in the inner subplot Top 10 Categories. For example, there were about 500 publications grouped in the 1st and 2nd clusters of the top 10, i.e., Cluster 188 and Cluster 35. In comparison, the 3rd to 6th clusters contained about 330-430 publications and less than 300 publications per each of the remaining clusters. Moreover, the Top 10 Clusters included about 20% of publications' distribution. It is worth mentioning that the clustering method proposed in this work was not designed to categorize a single publication in more than one category. Therefore, there was no multidisciplinary distribution of publications.

Clustering Quality Test Result

The proposed clustering method in this work was tested on a non-public dataset that suffered from the absence of ground-truth labels; this was because of the lack of such studies in this field. Hence, this case complicated the evaluation of the performance of the clustering method and the proposed profile matching approach. However, the presented "K-value determination" subsection showed that the performance of the clustering method at K=275 is the best among the tested values of K, as well as the statistical Kolmogorov Smirnov test at that value of K, i.e., 275, showed that the distribution of samples among the clusters fits the normal distribution with a value of $p < 0.05$. These results indicated the performance quality of the proposed clustering method. Moreover, the correlation between the resulted clusters was tested. Figure 14 shows a heatmap diagram that visualized the correlation analysis among the identified clusters.

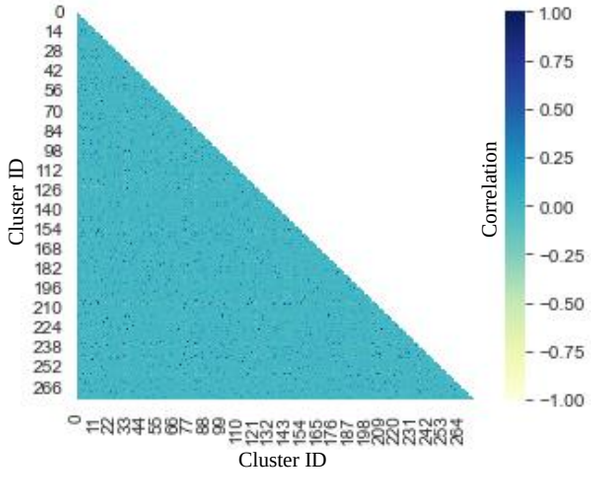


Figure 14 Visualization of Correlation between The Generated Clusters

Figure 14 showed that very few clusters were highly correlated dots in dark blue color. In contrast, the correlation between the majority of clusters ranged between -0.25 – 0.25, which indicated a good separation between clusters.

Profile matching Results

In addition to the clustering process, the proposed method aimed to match the researcher profiles through correlation-based similarity. For each identified cluster, the matrix of correlation between all researchers' publications within the cluster was calculated. The top similar publications were selected, and the researchers were proposed to be the best matching profiles of the selected researcher. As a result of this process, each researcher would be associated with some other researchers based on the similarity of their publication. Figure 15 shows a sample of the results of this process. Table 4 shows the description of the columns in Figure 15 starting from the left, which illustrates the output of the proposed method.

ResearcherProfileID	Group	Rec. Res. ProfileID	Most Similar Publication ID	Similarity	Publication Title
ec72EnIAAAAJ	5	5bnkmZUAAAAJ	5bnkmZUAAAAJ:t7zJ5fGR-2UC	0.899	changing perception harm e cigarette among u adult public health chronic disease tobacco use
		5bnkmZUAAAAJ	5bnkmZUAAAAJ:LO7wyVUgiFcC	0.899	changing perception harm e cigarette v cigarette use among adult u national survey public healtl
		vP6PjG4AAAAJ	vP6PjG4AAAAJ:a0OBvERweLwC	0.856	changing perception harm e cigarette among u adult tobacco global health
		0iqZYjcAAAAJ	0iqZYjcAAAAJ:e5wmG9Sq2KIC	0.310	relationship chronic lung disease status e cigarette use potential influence excessive alcohol use
		0iqZYjcAAAAJ	0iqZYjcAAAAJ:4JMBOYKVnBMC	0.310	dual use electronic cigarette traditional cigarette among adult psychosocial correlate associated r
ec72EnIAAAAJ	123	nmIR-egAAAAJ	nmIR-egAAAAJ:9vf0nzSNQJEC	0.894	benefit working informant criminology sociology
		nmIR-egAAAAJ	nmIR-egAAAAJ:N5tVd3kTz84C	0.894	working informant criminology sociology
		iirGxtEAAAAJ	iirGxtEAAAAJ:dshw04ExmUIC	0.913	protein secondary structure analysis dried blood serum using infrared spectroscopy identify mark
		JfxpzOQAAAAJ	JfxpzOQAAAAJ:efIP2zaiRacC	0.953	response drug resistant epilepsy adult outcome trajectory failure two medication biostatistics
		JfxpzOQAAAAJ	JfxpzOQAAAAJ:08ZZubdj9fEC	0.953	drug resistant epilepsy adult outcome trajectory failure two medication biostatistics
ec72EnIAAAAJ	82	5bnkmZUAAAAJ	5bnkmZUAAAAJ:b1wdh0AR-JQC	0.899	motif perception regarding electronic nicotine delivery system end use among adult mental healt
		vP6PjG4AAAAJ	vP6PjG4AAAAJ:t7zJ5fGR-2UC	0.896	use electronic nicotine delivery system end among chinese adult evidence citywide representati
		vP6PjG4AAAAJ	vP6PjG4AAAAJ:_axFR9aDTf0C	0.896	use electronic nicotine delivery system end china evidence citywide representative survey five cl
		SoO1Xm0AAAAJ	SoO1Xm0AAAAJ:UeHWp8X0CEIC	0.826	motif perception regarding electronic nicotine delivery system end use among adult mental healt
		SoO1Xm0AAAAJ	SoO1Xm0AAAAJ:KlAtU1dfN6UC	0.826	use electronic nicotine delivery system among adult mental health condition
ec72EnIAAAAJ	252	cNsXS68AAAAJ	cNsXS68AAAAJ:isC4tDSrTZIC	0.972	sankofa go back fetch merging genealogy africana study introduction literature humanity african a
		cNsXS68AAAAJ	cNsXS68AAAAJ:RGFaLdJalmkC	0.972	sankofa go back fetch merging genealogy africana study literature humanity african american stud
		SIhUdiEAAAAJ	SIhUdiEAAAAJ:roLk4NBRz8UC	0.806	intercultural communication education sla intercultural competence study abroad culture hyperm
		JfxpzOQAAAAJ	JfxpzOQAAAAJ:kuK5TVdYjLIC	0.214	study statistic knowledge among nurse faculty school research doctorate program biostatistics
		JfxpzOQAAAAJ	JfxpzOQAAAAJ:Y5dfb0dijaUC	0.214	mixed method study contraceptive effectiveness relationship context among young adult primari
EHSVmZ8AAAAJ	2	hEqmxx8AAAAJ	hEqmxx8AAAAJ:MLfJN-KU85MC	0.945	reciprocal relationship depressive symptom employment status labor economics demography ho
		hEqmxx8AAAAJ	hEqmxx8AAAAJ:z_wVstp3MssC	0.945	bilateral relationship depressive symptom employment status labor economics demography hou

Figure 15 Sample of Profile Matching Results

Table 4 Description of Output Data

Column Header	Description
ResearcherProfileID	The researcher Profile ID on GS.
Group	The Category ID(s) (i.e., Research Fields) as detected by the clustering method, note that some researchers are identified to be working in multiple research fields
Rec. Res. ProfileID	The researcher profile IDs whom were detected as top matched researches by the method.
Most Similar Publication ID	The GS id of the publication that belongs to the matched users.
Similarity	The similarity value (correlation) between the identified publication and the publications of the researches.
Publication Title	The publication title form GS.

Figure 15 showed that the proposed method was able to identify the top matched profiles of a Researcher based on the textual analysis of publication titles included in researchers' profiles on GS. The output showed that the publication titles in each group were similar as they had several common words, which were indeed similar to some publications in the Researcher profile under inspection. Additionally, some researchers were categorized in multiple categories where

each category, i.e., groups included similar publications from various user profiles. Furthermore, some identified publications had low similarity values in some groups marked in red color in Figure 15. However, some threshold cut value could be set for such cases to exclude such publication from the group if needed.

RESEARCH CONTRIBUTIONS

This research contributed to the domain by the following:

- The employment of the Unsupervised Machine Learning for solving the Researcher Profiles clustering problem.
- The employment of the correlation-based similarity for solving the Researcher Profiles matching problem.
- The analysis of results revealed hidden information about the scholarly work represented in the considered dataset. However, any institution could reveal such information using the same methods and analysis

CONCLUSIONS

This research aimed to solve the problem of profile matching in Scientific Research and Scholarly Work by employing unsupervised machine learning methods. The Vector Space Model (VSM) based on the term count vectorization and the PCA feature reduction methods were used to represent the data for the proposed machine learning method. Then, the K-mean clustering method was utilized to carry out the task of grouping or clustering the researcher profiles based on the statistical analysis of publication titles of the researchers. The correlation-based similarity was employed for profile matching within the clusters. The method was tested on an extracted dataset from Google Scholar. After preprocessing and filtering, the dataset contains the publication titles of 19866 publications which belong to 882 researchers from Georgia State University (GSU). The publications were categorized into 275 categories, i.e., Research Fields based on the analysis of clustering quality measures Distortion, Davies-Bouldin Index, Calinski-Harabasz Index, and the Kolmogorov Smirnov test. The proposed methods were implemented in python, and the analysis of the results revealed statistical information about the dataset. Moreover, the profile matching results and the clustering quality test result showed that the proposed method accomplished the designed task with high similarity of publications within the clusters and low correlation values among the clusters. The future direction of the research in this field included but was not limited to working on multi-lingual and larger datasets, testing various weighting methods, unsupervised machine learning, quality performance measures or studying the effect of dataset size and quality results generalization.

References

- *Andrews, N. O., and Fox, E. A. (2007). Recent Developments in Document Clustering: Department of Computer Science, Virginia Polytechnic Institute & State ….*
- *Deelers, S., and Auwatanamongkol, S. J. I. J. o. C. S. (2007). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning Along the Data Axis with the Highest Variance. 2(4): 247-252.*
- *Delua, J. (2021). Supervised Vs. Unsupervised Learning: What's the Difference? Artificial intelligence Retrieved 05/09/2021, 2021, from https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning*
- *Erisoglu, M., Calis, N., and Sakallioglu, S. (2011). A New Algorithm for Initial Cluster Centers in K-Means Algorithm. Pattern Recognition Letters. 32(14): 1701-1705.*
- *Eze, B., Kuziemsky, C., and Peyton, L. (2020). A Configurable Identity Matching Algorithm for Community Care Management. Journal of Ambient Intelligence and Humanized Computing. 11(3): 1007-1020.*
- *Feldman, R., and Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press.*
- *Franklin, J. (2005). The Elements of Statistical Learning: Data Mining, Inference and Prediction. The Mathematical Intelligencer. 27(2): 83-85.*
- *Garbade, M. J. (2018). Understanding K-Means Clustering in Machine Learning. Towards Data Science Retrieved 05/09/2021, 2021, from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1*
- *Garcia, P. E. (2016). Hybrid Algorithm for Matching Profiles and Social Networks.*
- *Jain, A. K., Murty, M. N., and Flynn, P. J. J. A. c. s. (1999). Data Clustering: A Review. 31(3): 264-323.*
- *Li, S., Lv, X., Wang, T., and Shi, S. (2010). The Key Technology of Topic Detection Based on K-Means. 2010 International Conference on Future Information Technology and Management Engineering. 387-390.*
- *Li, Y., Peng, Y., Zhang, Z., Yin, H., and Xu, Q. (2019). Matching User Accounts across Social Networks Based on Username and Display Name. World Wide Web. 22(3): 1075-1097.*
- *Milojević, S. (2013). Accuracy of Simple, Initials-Based Methods for Author Name Disambiguation. Journal of Informetrics. 7(4): 767-773.*

- *Milojević, S. (2014). Principles of Scientific Research Team Formation and Evolution. Proceedings of the National Academy of Sciences. 111(11): 3984-3989.*
- *Nurgaliev, I., Qu, Q., Bamakan, S. M. H., and Muzammal, M. (2020). Matching User Identities across Social Networks with Limited Profile Data. Frontiers of Computer Science. 14(6): 146809.*
- *Paembonan, S., Manga, A. R., Jusmidah, Atmajaya, D., Waluyantari, A. V., Astuti, W., and Mansyur, S. H. (2018). Combination of K-Means and Profile Matching for Drag Substitution. 2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT). 6-7 Nov. 2018. 180-183.*
- *Petrovic, S. (2006). A Comparison between the Silhouette Index and the Davies-Bouldin Index in Labelling Ids Clusters. Proceedings of the 11th Nordic Workshop of Secure IT Systems. 53-64.*
- *Pfitzner, D., Leibbrandt, R., and Powers, D. (2008). Characterization and Evaluation of Similarity Measures for Pairs of Clusterings. Knowledge and Information Systems. 19(3): 361.*
- *Pizzi, C., and Ukkonen, E. (2008). Fast Profile Matching Algorithms — a Survey. Theoretical Computer Science. 395(2): 137-157.*
- *Ray, S., and Turi, R. H. (1999). Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation. Proceedings of the 4th international conference on advances in pattern recognition and digital techniques. 137-143.*
- *Redmond, S. J., and Heneghan, C. (2007). A Method for Initialising the K-Means Clustering Algorithm Using Kd-Trees. Pattern Recognition Letters. 28(8): 965-973.*
- *Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., and Fujita, H. (2017). Modified Frequency-Based Term Weighting Schemes for Text Classification. Applied Soft Computing. 58: 193-206.*
- *Santos, R. S., Malheiros, S. M. F., Cavalheiro, S., and de Oliveira, J. M. P. (2013). A Data Mining System for Providing Analytical Information on Brain Tumors to Public Health Decision Makers. Computer Methods and Programs in Biomedicine. 109(3): 269-282.*
- *Sharma, S., and Gupta, V. J. I. J. o. C. A. (2012). Recent Developments in Text Clustering Techniques. 37(6): 14-19.*
- *Sugiarto, I., Diyasa, G. S. M., and Idhom, M. (2021). Profile Matching Algorithm in Determining the Position of Colleagues. Journal of Physics: Conference Series. 1844(1): 012026.*
- *Sun, C., Wan, Y., and Chen, Y. (2009). Dynamics of Research Team Formation in Complex Networks. Complex Sciences. 2009//. Berlin, Heidelberg. 2004-2015.*
- *Tran, N.-Y., Chan, E. K. J. C., and Libraries, R. (2020). Seeking and Finding Research Collaborators: An Exploratory Study of Librarian Motivations, Strategies, and Success Rates. 81(7): 1095.*
- *Wang, X., and Xu, Y. (2019). An Improved Index for Clustering Validation Based on Silhouette Index and Calinski-Harabasz Index. IOP Conference Series: Materials Science and Engineering. 569: 052024.*
- *Wassermann, B., and Zimmermann, G. (2011). User Profile Matching: A Statistical Approach. CENTRIC 2011, The fourth international conference on advances in human-oriented and personalized mechanisms, technologies, and services. 60-63.*
- *Wilcox, R. (2017). Comparing Two Groups. In: R. Wilcox (ed.). Introduction to Robust Estimation and Hypothesis Testing (Fourth Edition) (pp. 145-234): Academic Press.*
- *Yuan, C., and Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. 2(2): 226-235.*
- *Zhang, D., and Li, S. (2011). Topic Detection Based on K-Means. 2011 International Conference on Electronics, Communications and Control (ICECC). 2983-2985.*